TASKS, DOMAINS, AND LANGUAGES FOR INFORMATION EXTRACTION

Boyan Onyshkevych
Mary Ellen Okurowski
Lynn Carlson

U. S. Department of Defense
Ft. Meade, MD 20755
email: {lmcarls, baonysh, meokuro}@afterlife.ncsc.mil

1. TASKS

The information extraction tasks for the ARPA TIPSTER program center on automatically filling object-oriented data structures, called *templates*, with information extracted from free text in news stories (for discussion of templates and objects, see "Template Design for Information Extraction" in this volume). With text as input, the TIPSTER systems first detect whether the text contains relevant information. If so, the systems extract specific instances of generic types of information that correspond to each slot in the template and output that information by filling the template slots in an appropriate data representation. These slots are then scored by using an automatic scoring program with templates produced by human analysts that serve as answer keys. Human analysts also prepared development set templates for each domain, which served as training models for system developers (for discussion of the data preparation effort, see "Corpora and Data Preparation for Information Extraction" in this volume).

With the TIPSTER program goal of demonstrating domain and language-independent algorithms, extraction tasks for two domains (joint ventures and microelectronics chip fabrication) for both English and Japanese were identified. The selection criteria for this pair of languages included linguistic diversity, availability of on-line resources, and availability of computer support resources. The four pairs include EJV, JJV, EME, and JME, abbreviated to reflect the language (E or J) and the domain (JV or ME). The tasks, domains and languages used for the information extraction portion of the TIPSTER program were also used for the Fifth Message Understanding Conference (MUC-5). In MUC-5, non-TIPSTER participants could choose to perform in one of the domains in Japanese and/or English. Of the TIPSTER participants, three performed in all four pairs, and the fourth in both domains but only in English.

2. THE JOINT VENTURE DOMAIN

The reporting task for the domain of Joint Ventures involves capturing information about business activities of entities (companies, governments, individuals, or other organizations) who enter into a cooperative agreement for a specific project or purpose. The partnership formed between these entities may or may not result in the creation of a separate joint venture company to carry out the activities of the agreement. In many cases, a looser cooperative arrangement between partner entities (called a 'tie up') is established; information about tie ups is also captured as part of the JV task. A tie up may involve joint product development, market share arrangements, technology transfer, etc. The terms 'tie up' and 'joint venture' are sometimes used interchangeably in describing the Joint Venture domain. In addition to reporting information about new joint ventures, the JV task also involves capturing information from reports of the activity of existing joint ventures or about changes in any joint venture agreements. In other words, any discussion of joint ventures in any news article is to be reported, so long as enough information is presented to meet the minimal reporting conditions, namely, that at least two entities are involved in a two-way discussion about forming a joint venture, or have entered into such an agreement, and that at least one piece of identifying information is known about each involved entity, such as its name, nationality, or location.

The JV template consists of eleven different object types, which together capture essential information about joint venture formation and activities (the canonical *who*, *what*, *where*, *when*, and *why*). The `TIE-UP-RELATIONSHIP` object captures the most basic information about a tie up or joint venture discussed in a particular document, including who the tie-up partner `ENTITY`s are, and who the joint venture (or "child") company is, if one is formed. Additionally, for the `TIE-UP-RELATIONSHIP`, the `STATUS` of the joint venture is recorded, along with information about `OWNERSHIP` and `ACTIVITY`s associated with the tie up. The `ACTIVITY` involves information

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
SEP 1993		**00-00-1993 to 00-00-1993**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Tasks, Domains, and Languages for Information Extraction	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **U. S. Department of Defense,Fort Meade,MD,20755**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
TIPSTER TEXT PROGRAM: PHASE I: Proceedings of a Workshop held at Fredericksburg, Virginia, September 19-23, 1993. Sponsored by the Advanced Research Projects Agency.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**11**	

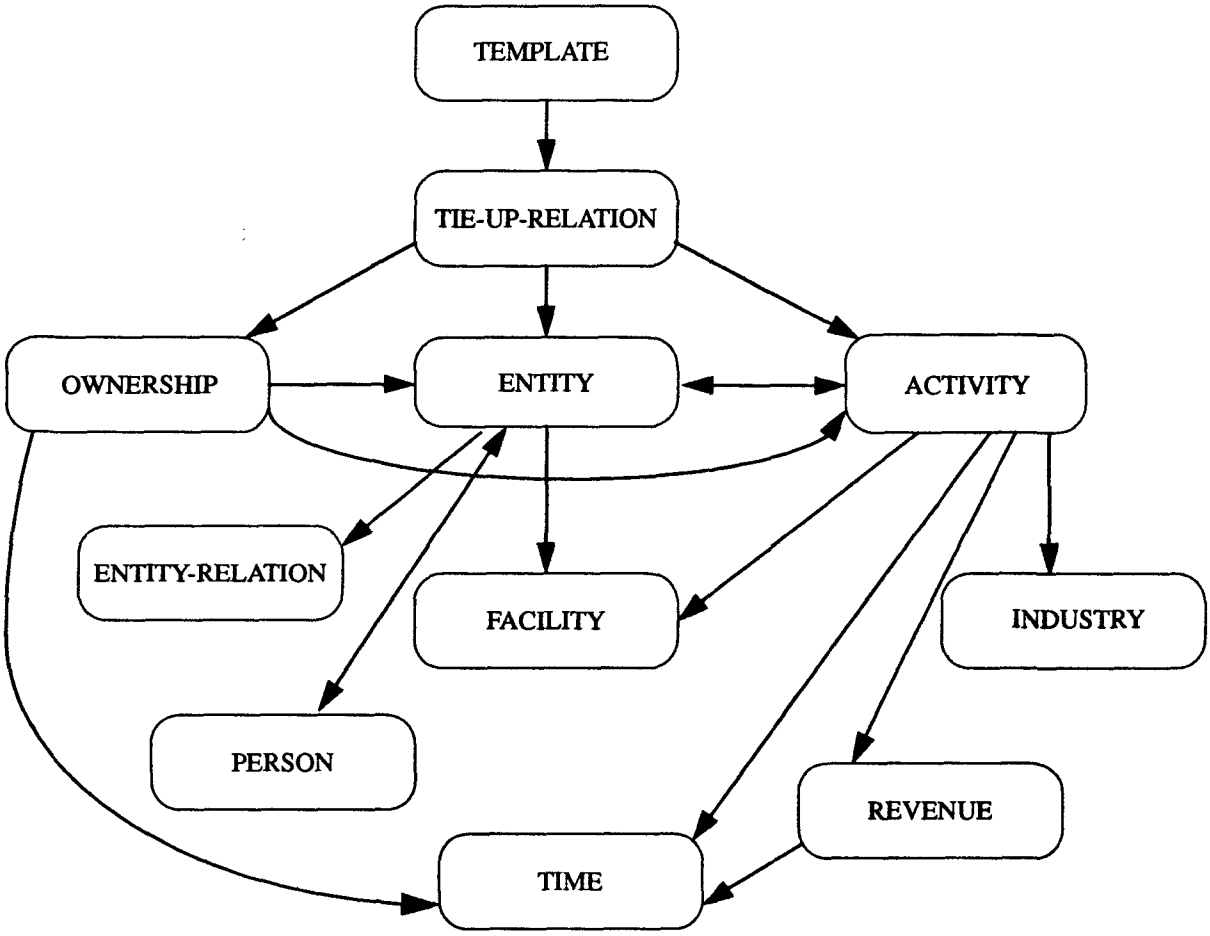

Figure 1: Joint Venture template object types (and pointers)

about what the joint venture will be doing, including what INDUSTRYs the tie up will be involved in, where the activity will take place, when it will start and end, and how much REVENUE is expected from the venture. The INDUSTRY of the joint venture is captured in categorical terms (MANUFACTURING, RESEARCH, SALES, etc.), and also coded as one or more Standard Industrial Classification categories (see Section 4 below) linked with the specific words in the text that define the business.

Figure 1 illustrates the object types and the interrelations among them in the Joint Ventures domain. Note that multiple interrelations may be represented by one arrow; for example, the TIE-UP-RELATIONSHIP object has two possible interrelationships with ENTITY objects, i.e., either identification of the parents in a tie up, or identification of the joint venture child company itself. The relative complexity of the template design mirrors the intricacies inherent in the tie-up event itself.

Appendix A gives a straightforward example from the English Joint Venture domain, including an excerpt from an EJV article, along with its corresponding filled-out template. There is one tie up in this article, triggered by the announcement that "Bridgestone Sports Co. ... has set up a joint venture in Taiwan." This tie up has three parent or partner companies ("Bridgestone Sports Co.," "Union Precision Casting Co.," and "Taga Co.") and a child company ("Bridgestone Sports Taiwan Co.") that will be engaged in the production of golf clubs. Although there is only one TIE-UP in this article, multiple tie ups in a single article are common in the English and Japanese corpora. In the diagram in Figure 3, note the different labels on the arcs; e.g., the TIE-UP-RELATIONSHIP has two types of arcs pointing to ENTITYs, reflecting the two types of interrelationships discussed above, namely, "parent company" and "joint venture child company."

In Appendix C, an example is given from the Japanese

Joint Venture domain. The excerpted article references a new activity about to get underway in the financial arena, namely, issuing of a new credit card. This new product is the result of a recent tie up between "Daimaru" and "six companies of the VISA Card Group, including "Sumitomo Credit Service." In accordance with the JJV fill rules (see "Corpora and Data Preparation for Information Extraction" in this volume), the tie up is instantiated between Daimaru and Sumitomo Credit Service, which is regarded as the group leader for the VISA Card Group, because it is the only group member explicitly mentioned in the text. The template indicates a single ACTIVITY, with an INDUSTRY type FINANCE, and product/service string "issuing the Daimaru Excel VISA card."

3. THE MICROELECTRONICS DOMAIN

The reporting task in the domain of Microelectronics involves capturing information about advances in four types of chip fabrication processing technologies: layering, lithography, etching, and packaging. For each process, this information relates to process-specific parameters that typify advancements. For example, the introduction of a new type of film used in layering (that is, adding material to the substrate of a chip) or an increase in the resolution possible in lithography both indicate new developments in fabrication technology. To be reported, each process must be associated with some identifiable entity that is manufacturing, selling, or distributing equipment, or developing or using processing technology.

The MICROELECTRONICS-CAPABILITY template object links together information about the four fabrication technologies (LITHOGRAPHY, LAYERING, PACKAGING, and ETCHING) with the ENTITYs, typically companies, associated with one of the technologies as its DEVELOPER, MANUFACTURER, DISTRIBUTOR, or PURCHASER_-USER. Additionally, the template captures information about the specific EQUIPMENT used, developed, or sold, as well as information about the type of chips or DEVICEs that are expected to be produced by that technology. There are nine objects in the domain.

Figure 2 illustrates the object types captured in the Microelectronics template. Appendix B provides an example from the Microelectronics domain, including an excerpt from an EME article, along with its corresponding filled-out template. There are two microelectronic capabilities in this example. The first capability is succinctly represented in the first sentence with the identification of a lithography process ("a new stepper") associated with an entity ("Nikon Corp.") as the manufacturer and distributor ("to market") of a piece of equipment that implements a lithographic process. Note also that the technology will be used to produce a device ("64-Mbit DRAMs"), which satisfies the reporting condition requirement for technology connection to integrated circuit production. Additional information about the process and equipment occurs in the text. The second capability stems from information in the second sentence (i.e., "compared to the 0.5 micron of the company's latest stepper"). The need to interpret this segment within the context of the discourse demonstrates the level of text understanding required in this domain.

4. DOMAIN DIFFERENCES

The JV and ME domains differ in the focus of their task, type of complexity, and level of technicality. The focus of the JV task is the tie-up formation and the corresponding activities of the resulting agreement. Thus, to a large extent, the task is event-driven. The information to be extracted includes the participants in the event, the economic activity of the event, and adjunct information about the event, such as time, facilities, revenue, and ownership. Entities are central, specifically within the context of the tie-up relationship. In addition, relationships also dominate in that the tie-up event presents a cohesive collection of linked objects, e.g., persons and facilities linked to entities, entities linked to other entities, industries linked to activities, and so on. The overarching task is fitting together the interrelated pieces of the single tie-up event.

The focus of the ME task is the four microelectronics chip fabrication processes and their attributes. The task is not triggered by a particular event, as in JV; the focus is on more static information. The information to be extracted includes the processes with their attributes and associated devices and pieces of equipment. Processes are central in ME, whereas entities are in some sense auxiliary. Although clearly the information about processes must be associated with an entity to be relevant, the task design centers on the processes themselves and their attributes. Essentially, the domain fractures into four separate sub-tasks, one for each process. Linking attributes to a process, like film or temperature to the layering process, involves defining the process in terms of key characteristics inherent in the process itself. Both devices and pieces of equipment are also associated with processes, but in quite different and indirect ways. Equipment represents the *implementation* of a process, whereas devices represent the *application* of a process. No single overarching task applies for the ME domain; rather, there are four separate, concurrent subtasks in which associated characteristics of processes are identified.

The two domains also differ in the nature of their complexity. The complexity of the JV domain lies not in the predominance of technical jargon but in the intricacies of the interrelationships within a tie-up event. These intricacies

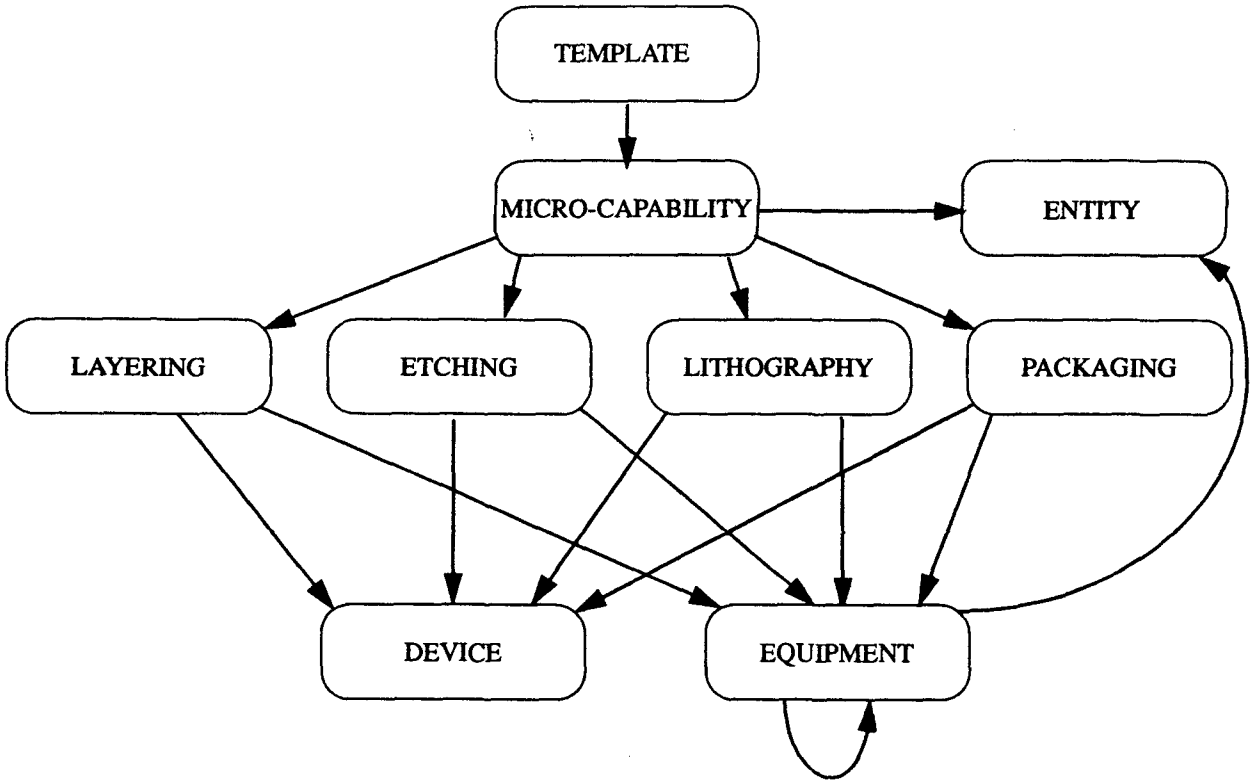

Figure 2: Microelectronics template object types (and pointers)

cover a broad range of activities that legitimately fall within the domain of joint business ventures. Since there is no single way to create a business relationship of the sort captured in this domain, there can be many points at which interpretation or judgment comes into play. Although this interpretation can be minimized by specification (sometimes arbitrary) in the fill rules, the open-endedness, and in some ways potential for creativity, in how a tie-up is realized results in domain complexity. For example, determining whether or not a text has enough information to warrant reporting a tie up, or whether there is sufficient evidence for a tie-up activity, may require a substantial amount of judgment on the part of the analyst. Initially, there was a wide variation in interpretation of these issues among the JV analysts for each language. However, through frequent meetings, these differences in interpretation were narrowed over time, and there was a convergence of viewpoints on what information to extract from a given JV document and how to represent it in the template. The fill rules were continually modified and updated to incorporate the heuristics developed by the analysts for determining when a valid tie up or activity existed.

The resolution of coreferences, which also contributes to domain complexity, is a key task in the Joint Ventures domain. In particular, the entities in the JV documents were typically referenced in multiple ways. The EJV example in Appendix A illustrates one case where each of the ENTITYs is referred to at least three times in the text, and each of those multiple (and differing) references may contribute additional information to the ENTITY objects. For example, the phrase "the Japanese sports goods maker" needs to be coreferenced with "Bridgestone Sports Co." in order to identify the nationality of Bridgestone. Of equal importance in the JV domain is event-level coreference determination, in other words, determining which joint ventures are unique among a set of multiple apparent joint ventures in the text. For example, the article in Appendix A has multiple paragraphs, each discussing a joint venture, and event-level coreference resolution is required to determine that they are

all discussing the same joint venture, not four different ones. This coreference layering problem at both entity and event levels makes extraction difficult in this domain.

In comparison, the ME domain derives complexity not from interrelationships, but from its composition. There are four sub-domains, one for each process. Each sub-domain corresponds to a process with attributes, two of which can be devices or pieces of equipment. In addition, entities are associated with these processes in one of four different capacities: developer, manufacturer, distributor, or purchaser/user. Adding complexity to the ME domain is the pre-requisite to connect the technology to integrated chip production.

The third area of domain difference is the level of technicality, namely, the extent to which highly technical terms and knowledge are used. The JV domain lies within the financial/economic area, and the articles are typical of general business news. The one element of the JV domain that relies more on technical jargon or specific technical descriptions is the product or service that the joint venture will be involved in. This information, in addition to being reported as an exact string fill from the text, also is reported in the JV template as a two-digit code, according to the Standard Industrial Classification manual of the U. S. Office of Management and Budget. These strings may involve technical terms; for example, "ignition wiring harness" is classified under SIC 36, which includes electronic and other electrical equipment and components.

In contrast, the ME domain lies within the scientific and technical arena with a corpus composed of product announcements and reports on research advances. The texts are loaded with domain-specific technical terms, at times detailing chip fabrication methodology. The fill rules provide a resource for this technical terminology, which essentially provides hooks into the text for extracted information. These hooks mean that in the pre-processing stage, some of the extracted information can be identified as discrete tagged elements and then confirmed for extraction in later stages of processing. This "bias for keywording" is lessened to some extent by the higher percentage of irrelevant documents in the ME corpus than the JV corpus and by two requirements in the reporting conditions (i.e., a process must be associated with an entity in one of four roles and the application for the process must be related to integrated circuits).

5. LANGUAGE DIFFERENCES

Although the Japanese and English tasks are apparently identical (other than the language of the texts and templates), subtle differences emerge with closer scrutiny of the corpora, template definitions, and fill rules (see "Corpora and Data Preparation for Information Extraction" in this volume) for each of the two languages. The corpora for English and Japanese differ, in that the two English corpora are drawn from more than 200 sources each, and have a fairly low percentage of irrelevant documents in the set, whereas the Japanese corpora have a limited set of sources, but a higher percentage of irrelevant documents.

Over the course of the data preparation task, differences between the English and Japanese texts were gradually identified and strategies for dealing with them were incorporated into the fill rules. A major difference between texts in the JV domain is the fact that in the JJV corpus, the most typical relationship involves two entities joining together in a tie up where no joint venture company is created, whereas in EJV, the typical relationship involves one in which two entities form a joint venture company as part of the agreement. In EJV, texts which were produced by Japanese news sources (in English) could also reflect the type of tie-up arrangement typical of the Japanese texts, i.e., where no joint venture company is formed.

Differences between Japanese and English are also reflected in minor discrepancies in the Japanese and English template definitions and more substantial divergences in the corresponding fill rules. While every attempt was made to keep the template definition for each domain identical across languages, there are some differences. Thus, although the English and Japanese templates have the same objects and slots for each domain, there are cases where the content or format of the fills for a particular slot vary from one language to the other, reflecting differences in the two corpora.

In the JJV and EJV templates, an example of a content difference in fillers is seen in the FACILITY object's FACILITY-TYPE slot, which is a set fill for both EJV and JJV. However, for EJV the fillers include COMMUNICATIONS, SITE, FACTORY, FARM, OFFICE, MINE, STORE, TRANSPORTATION, UTILITIES, WAREHOUSE, and OTHER, whereas in JJV, the fillers are (translated): STORE, RESEARCH_INSTITUTE, FACTORY, CENTER, OFFICE, TRANSPORTATION, COMMUNICATIONS, CULTURE/LEISURE, and OTHER. The fillers were defined and selected by the analysts to reflect the types of information most commonly found in the corpora.

A format difference in slot fills between languages (for both JV and ME) is exhibited in the ENTITY object's NAME slot, where English requires a normalized form for the entity name, based on a standardized list of abbreviations for corporate designators, including more familiar ones like INC (incorporated) and LTD (limited), as well as some spe-

cifically used by foreign firms, such as AG (for Aktiengesellschaft -- Germany), EC (for Exempt Company -- Bahrain), and PERJAN (for Perusuhan Jawatan -- Indonesia). For Japanese, such a list of designators was not available, and in the corpus itself, most companies are indicated by the ending *sha* or *kaisha*, so it was decided that a string fill would be more appropriate for this slot filler.

The JJV fill rules give detailed decision trees for determining who the tie-up partners are. This reflects the fact that in the JJV corpus, the texts often begin by mentioning a tie up between two groups. For example, the two groups might be Mitsubishi Group and Daimler Group, but then, in the second paragraph, one learns that the actual tie up is between Mitsubishi Shoji and Daimler Benz. The JJV fill rules explicitly address this type of situation, since it occurs frequently in the corpus. The fill rules stipulate that in cases of tie ups between groups, the group leaders are to be taken as the tie-up partners, if they are mentioned in the text. The EJV fill rules address a slightly different problem, namely, of how to represent tie-up partners if the text states "Four Malaysian finance firms announced a joint venture...," in which case a tie up between two (not four) identical partner entities would be created. This situation did not typically arise in the JJV corpus.

As with the JV domain, the two ME corpora highlight significant differences. First, there are basically three news sources for the JME corpus, the same set of sources as for the JJV corpus. The EME corpus, on the other hand, is selected from a business and trade database with more than 200 different sources. Second, the JME corpus (30%) contains a higher percentage of irrelevant documents than the English corpus (20%). Third, even though the relative proportion of the four process types is similar, there is a distinct difference between languages in the type of information available for the PACKAGING object. Not only is there considerably more information available for all PACKAGING slots for English, there is also clear evidence that information for the BONDING and UNITS_PER_-PACKAGE slots is infrequently available in Japanese. English texts are also more likely than Japanese texts to contain two or more PACKAGING objects, which may partially explain anecdotal reports that PACKAGING texts were considered difficult to code for the English analysts, but easy for the Japanese analyst.

There are actually no substantive differences reflected in the two sets of fill rules for the ME domain. However, differences between languages are indicated in the type of information available for extraction. For example, some set fill choices in the template simply do not occur in Japanese, like some of the hierarchical set fill choices for the PACKAGING object's TYPE slot in ME. Also the keywords, "gate size" and "feature size" that indicate granularity for the LITHOGRAPHY object do not occur in the Japanese corpus. Other minor differences are also indicated in the fill rules as to how the information is represented in the English and Japanese texts. To illustrate, in contrast to the EME fill rules, the Japanese fill rules are more likely to list relevant keywords in the text associated with ENTITY roles and to identify relevant stereotypic format clues for location information. This approach suggests the greater likelihood of identifiable patterns within the Japanese text. Another illustration of the dissimilarity in information presentation is the Japanese inclusion of English within the Japanese text, for example in layering or packaging types or in entity names.

APPENDIX A: Example from English Joint Ventures

Source document sections: (Note the SGML tags delimiting document and headers.)

```
<doc>
<DOCNO>  0592 </DOCNO>
<DD>     NOVEMBER 24, 1989, FRIDAY  </DD>
<SO>     Copyright (c) 1989 Jiji Press Ltd.</SO>
<TXT>

      BRIDGESTONE SPORTS CO. SAID FRIDAY IT HAS SET UP A  JOINT VENTURE  IN TAIWAN  WITH A
LOCAL CONCERN AND A  JAPANESE  TRADING HOUSE TO PRODUCE GOLF CLUBS TO BE SHIPPED TO  JAPAN.
      THE JOINT VENTURE, BRIDGESTONE SPORTS TAIWAN CO., CAPITALIZED AT 20 MILLION NEW TAIWAN
DOLLARS, WILL START PRODUCTION IN JANUARY 1990 WITH PRODUCTION OF 20,000 IRON AND "METAL WOOD"
CLUBS A MONTH. THE MONTHLY OUTPUT WILL BE LATER RAISED TO 50,000 UNITS, BRIDGESTON SPORTS
OFFICIALS SAID.
      THE NEW COMPANY, BASED IN KAOHSIUNG, SOUTHERN  TAIWAN,  IS OWNED 75 PCT BY BRIDGESTONE
SPORTS, 15 PCT BY UNION PRECISION CASTING CO. OF  TAIWAN  AND THE REMAINDER BY TAGA CO., A
COMPANY ACTIVE IN TRADING WITH  TAIWAN,  THE OFFICIALS SAID.
...
      WITH THE ESTABLISHMENT OF THE  TAIWAN  UNIT, THE  JAPANESE  SPORTS GOODS MAKER PLANS TO
INCREASE PRODUCTION OF LUXURY CLUBS IN  JAPAN.

</TXT>
</doc>
```

Template: (For explanation of notation, see "Template Design for Information Extraction" in this volume.)

```
<TEMPLATE-0592-1> :=
    DOC NR: 0592
    DOC DATE: 241189
    DOCUMENT SOURCE: "Jiji Press Ltd."
    CONTENT: <TIE_UP_RELATIONSHIP-0592-1>
    DATE TEMPLATE COMPLETED: 251192
<TIE_UP_RELATIONSHIP-0592-1> :=
    TIE-UP STATUS: EXISTING
    ENTITY: <ENTITY-0592-1>
            <ENTITY-0592-2>
            <ENTITY-0592-3>
    JOINT VENTURE CO: <ENTITY-0592-4>
    OWNERSHIP: <OWNERSHIP-0592-1>
    ACTIVITY: <ACTIVITY-0592-1>
<ENTITY-0592-1> :=
    NAME: BRIDGESTONE SPORTS CO
    ALIASES: "BRIDGESTONE SPORTS"
            "BRIDGESTON SPORTS"
    NATIONALITY: Japan (COUNTRY)
    TYPE: COMPANY
    ENTITY RELATIONSHIP: <ENTITY_RELATIONSHIP-0592-1>
<ENTITY-0592-2> :=
    NAME: UNION PRECISION CASTING CO
    ALIASES: "UNION PRECISION CASTING"
    LOCATION: Taiwan (COUNTRY)
    NATIONALITY: Taiwan (COUNTRY)
    TYPE: COMPANY
    ENTITY RELATIONSHIP: <ENTITY_RELATIONSHIP-0592-1>
<ENTITY-0592-3> :=
    NAME: TAGA CO
    NATIONALITY: Japan (COUNTRY)
```

```
    TYPE: COMPANY
    ENTITY RELATIONSHIP: <ENTITY_RELATIONSHIP-0592-1>
<ENTITY-0592-4> :=
    NAME: BRIDGESTONE SPORTS  TAIWAN  CO
    LOCATION: "KAOHSIUNG" (UNKNOWN) Taiwan (COUNTRY)
    TYPE: COMPANY
    ENTITY RELATIONSHIP: <ENTITY_REL-0592-1>
<INDUSTRY-0592-1> :=
    INDUSTRY-TYPE: PRODUCTION
    PRODUCT/SERVICE: (39 "20,000 IRON AND 'METAL WOOD' [CLUBS]")
             / (39 "GOLF [CLUBS]")
             / (39 "GOLF [CLUBS] TO BE SHIPPED TO  JAPAN")
<ENTITY_RELATIONSHIP-0592-1> :=
    ENTITY1: <ENTITY-0592-1>
            <ENTITY-0592-2>
            <ENTITY-0592-3>
    ENTITY2: <ENTITY-0592-4>
    REL OF ENTITY2 TO ENTITY1: CHILD
    STATUS: CURRENT
<ACTIVITY-0592-1> :=
    INDUSTRY: <INDUSTRY-0592-1>
    ACTIVITY-SITE: (Taiwan (COUNTRY) <ENTITY-0592-4>)
    START TIME: <TIME-0592-1>
<TIME-0592-1> :=
    DURING: 0190
<OWNERSHIP-0592-1> :=
    OWNED: <ENTITY-0592-4>
    TOTAL-CAPITALIZATION: 20000000 TWD
    OWNERSHIP-%: (<ENTITY-0592-3> 10)
                 (<ENTITY-0592-2> 15)
                 (<ENTITY-0592-1> 75)
```

Diagram: Figure 3 illustrates the template above in graphical form; notice the labels on some arcs, either indicating the slot in the object where the pointer resides, or (e.g., in OWNERSHIP), an associated value.

APPENDIX B: Example from English Microelectronics

Source document sections: (Note the SGML tags delimiting document and headers.)

```
<doc>
<REFNO>  000132038 </REFNO>
<DOCNO>  2789568 </DOCNO>
<DD>     October 19, 1990 </DD>
<SO>     Comline Electronics </SO>
<TXT>

In the second quarter of 1991, Nikon Corp. (7731) plans to market the "NSR-1755EX8A," a new
stepper intended for use in the production of 64-Mbit DRAMs.  The stepper will use an 248-nm
excimer laser as a light source and will have a resolution of 0.45 micron, compared to the 0.5
micron of the company's latest stepper.
...
      COMLINE NEWS SERVICE, Sugetsu Building, 3-12-7 Kita-Aoyama,Minato-Ku, Tokyo 107, Japan.
Telex 2428134 COMLN J.

</TXT>
</doc>
```

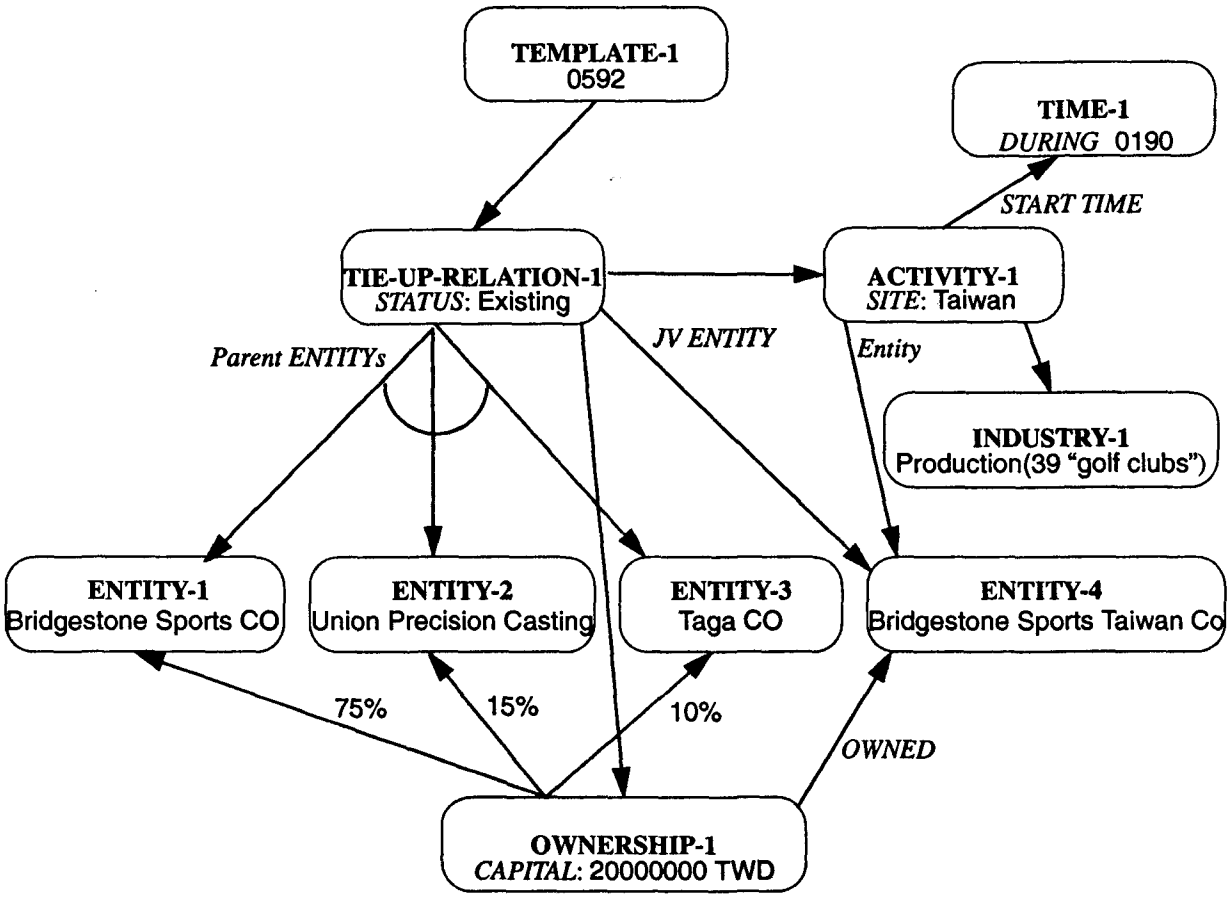

Figure 3: Diagram of (parts of) template for article 0592

Template: (For explanation of notation, see "Template Design for Information Extraction" in this volume.)

```
<TEMPLATE-2789568-1> :=
    DOC NR: 2789568
    DOC DATE: 191090
    DOCUMENT SOURCE: "Comline Electronics"
    CONTENT: <MICROELECTRONICS_CAPABILITY-2789568-1>
<MICROELECTRONICS_CAPABILITY-2789568-2>
    DATE TEMPLATE COMPLETED: 031292
    EXTRACTION TIME: 7
    COMMENT: / "TOOL_VERSION: LOCKE.5.2.0"
           / "FILLRULES_VERSION: EME.5.2.1"
<MICROELECTRONICS_CAPABILITY-2789568-1> :=
    PROCESS: <LITHOGRAPHY-2789568-1>
    MANUFACTURER: <ENTITY-2789568-1>
    DISTRIBUTOR: <ENTITY-2789568-1>
<MICROELECTRONICS_CAPABILITY-2789568-2> :=
    PROCESS: <LITHOGRAPHY-2789568-2>
    MANUFACTURER: <ENTITY-2789568-1>
```

```
 <ENTITY-2789568-1> :=
    NAME: Nikon CORP
    TYPE: COMPANY
<LITHOGRAPHY-2789568-1> :=
    TYPE: LASER
    GRANULARITY: ( RESOLUTION 0.45 MI )
    DEVICE: <DEVICE-2789568-1>
    EQUIPMENT: <EQUIPMENT-2789568-1>
<LITHOGRAPHY-2789568-2> :=
    TYPE: UNKNOWN
    GRANULARITY: ( RESOLUTION 0.5 MI )
    EQUIPMENT: <EQUIPMENT-2789568-2>
<DEVICE-2789568-1> :=
    FUNCTION: DRAM
    SIZE: ( 64 MBITS )
<EQUIPMENT-2789568-1> :=
    NAME_OR_MODEL: "NSR-1755EX8A"
    MANUFACTURER: <ENTITY-2789568-1>
    MODULES: <EQUIPMENT-2789568-3>
    EQUIPMENT_TYPE: STEPPER
    STATUS: IN_USE
<EQUIPMENT-2789568-2> :=
    MANUFACTURER: <ENTITY-2789568-1>
    EQUIPMENT_TYPE: STEPPER
    STATUS: IN_USE
<EQUIPMENT-2789568-3> :=
    MANUFACTURER: <ENTITY-2789568-1>
    EQUIPMENT_TYPE: RADIATION_SOURCE
    STATUS: IN_USE
```

APPENDIX C: Example from Japanese Joint Ventures

Source document sections: (Note the SGML tags delimiting document and headers.)

```
<doc>
<REFNO>  朝日新聞.000023 </REFNO>
<DOCNO>  0023 </DOCNO>
<DD>     85.03.12 </DD>
<SO>     朝日新聞　朝刊　８頁　２経　写図無　（全１７５字）　</SO>
<TXT>

　大丸は四月四日から、住友クレジットサービス（本社・大阪市）などＶＩＳＡカードグループ六社と提携した「大丸エクセルＶＩＳＡカード」を発行する。
...
大丸で買い物をした金額の５％が割引になるほか、ＶＩＳＡカードとして、国内や世界百六十五カ国の加盟店で通用する。

</TXT>
</doc>
```

Template: (For explanation of notation, see "Template Design for Information Extraction" in this volume.)

```
<テンプレート-0023-1> :=
    記事符号: 0023
    発行年月日: 850312
    ニュース出所: "朝日新聞　朝刊"
    内容: <提携-0023-1>
    完了年月日: 310393
    抽出時間: 15
    コメント: / "TOOL_VERSION: SHOTOKU.3.0.2"
              / "FILLRULES_VERSION: JJV.3.0.1"
<提携-0023-1> :=
    提携状況: 現行
    エンティティー: <エンティティー-0023-1>
                    <エンティティー-0023-2>
    経済活動: <経済活動-0023-1>
<エンティティー-0023-1> :=
    エンティティー名: 大丸
    エンティティー別: 企業
    エンティティー関係: <エンティティー関係-0023-1>
<エンティティー-0023-2> :=
    エンティティー名: 住友クレジットサービス
    場所: 日本(国)大阪府(県)大阪(市)
    エンティティー別: 企業
    エンティティー関係: <エンティティー関係-0023-1>
<業種-0023-1> :=
    業種別: 財政
    製品・サービス: (61 "[「大丸エクセルＶＩＳＡカード」を発行する]")
    コメント: "SIC 6153"
<エンティティー関係-0023-1> :=
    エンティティー乙: <エンティティー-0023-1>
                      <エンティティー-0023-2>
    甲対乙関係: パートナー
    状況: 現在
<経済活動-0023-1> :=
    業種: <業種-0023-1>
    場所: (- <エンティティー-0023-2>)
           (- <エンティティー-0023-1>)
```